Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/110777/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Monden, Rei, Roest, Annelieke M., van Ravenzwaaij, Don, Wagenmakers, Eric-Jan, Morey, Richard ORCID: https://orcid.org/0000-0001-9220-3179, Wardenaar, Klaas J. and de Jonge, Peter 2018. The comparative evidence basis for the efficacy of second-generation antidepressants in the treatment of depression in the US: A Bayesian meta-analysis of Food and Drug Administration reviews. Journal of Affective Disorders 235 , pp. 393-398. 10.1016/j.jad.2018.04.040 file

Publishers page: https://doi.org/10.1016/j.jad.2018.04.040 <https://doi.org/10.1016/j.jad.2018.04.040>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

THE COMPARATIVE EVIDENCE BASIS FOR THE EFFICACY OF SECOND-GENERATION ANTIDEPRESSANTS IN THE TREATMENT OF DEPRESSION IN THE US: A BAYESIAN META-ANALYSIS OF FOOD AND DRUG ADMINISTRATION REVIEWS

Rei Monden (PhD)[1], Annelieke M. Roest (PhD)[1,2], Don van Ravenzwaaij (PhD)[2], Eric-Jan Wagenmakers (PhD)[3], Richard Morey (PhD)[4], Klaas J. Wardenaar (PhD)[1], Peter de Jonge (PhD)[2]

[1] University of Groningen, University Medical Center Groningen, Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion regulation, Groningen, the Netherlands

[2] University of Groningen, Faculty of Behavioural and Social Sciences, Groningen, the Netherlands

[3] University of Amsterdam, Department of Psychology, Amsterdam, the Netherlands

[4] Cardiff University, School of Psychology, Cognitive Science, Wales, United Kingdom

Hanzeplein 1, P.O.Box 30.001 –CC72, 9700RB, Groningen, the Netherlands
+31 50 361 4490
Email address of the corresponding author:
RM: r.tendeiro-monden@umcg.nl

Abstract

Background Studies have shown similar efficacy of different antidepressants in the treatment of depression.

Method Data of phase-2 and -3 clinical-trials for 16 antidepressants (levomilnacipran, desvenlafaxine, duloxetine, venlafaxine, paroxetine, escitalopram, vortioxetine, mirtazapine, venlafaxine XR, sertraline, fluoxetine, citalopram, paroxetine CR, nefazodone, bupropion, vilazodone), approved by the FDA for the treatment of depression between 1987 and 2016, were extracted from the FDA reviews that were used to evaluate efficacy prior to marketing approval, which are less liable to reporting biases. Meta-analytic Bayes factors, which quantify the strength of evidence for efficacy, were calculated. In addition, posterior pooled effect-sizes were calculated and compared with classical estimations.

Results The resulted Bayes factors showed that the evidence load for efficacy varied strongly across antidepressants. However, all tested drugs except for bupropion and vilazodone showed strong evidence for their efficacy. The posterior effect-size distributions showed variation across antidepressants, with the highest pooled estimated effect size for venlafaxine followed by paroxetine, and the lowest for bupropion and vilazodone.

Limitations Not all published trials were included in the study.

Conclusions The results illustrate the importance of considering both the effect size and the evidence-load when judging the efficacy of a treatment. In doing so, the currently employed Bayesian approach provided clear insights on top of those gained with traditional approaches.

Key words

Food and Drug Administration (FDA); depression; antidepressant; Bayes factor; Bayesian Statistics

Introduction

Depression is one of the largest contributors to the global burden of disease (Murray et al., 2012; Ferrari et al., 2013; Whiteford et al., 2013; Compton et al., 2006). In 2013, it was estimated that 15.7 million adults in the US had at least one major depressive episode in the past year (National Institute of Mental Health, 2015). Given the large impact of depression on patients and society (Alonso et al., 2004; Kessler, 2012), implementing effective treatment is a key priority. Antidepressant medication is one of the most common treatments for depression with about a third of severely depressed patients using antidepressants in the US (Pratt et al., 2011). Partially because of their use for other disorders (e.g. anxiety disorders, chronic pain or insomnia) and the development of better tolerated Selective Serotonin Reuptake Inhibitors (SSRIs), the popularity of antidepressants has grown substantially: their consumption in the US increased by almost 400% between 1988-1994 and 2005-2008 (National Center for Health Statistics, 2010). The same trend was observed in other high income countries. For instance, in Germany, a 46% increase in antidepressants consumption was observed between 2007 and 2011 (OECD library, 2013).

An important aspect of the pharmacological treatment of depression is the selection of an antidepressant by the clinician. However, most antidepressants have been shown to have very similar, moderate effects in randomized controlled trials (RCT; Rush et al., 1995) and the APA's revised Practice Guideline for the Treatment of Major Depressive Disorder concludes that many antidepressants are equally effective (American Psychiatric Association, 2010;P.33). The similar efficacy of antidepressants in RCTs could be due to different antidepressant compounds acting on the brain in a similar way (National Institute for Health and Care Excellence, 2007), resulting in similar effects when compared to a placebo (i.e., effect sizes around 0.3; Rush et al., 2006; Turner et al., 2008; Kirsch et al., 2008). The lack of a clear differentiation between the efficacy of various antidepressants has led to a large variability in prescription behavior among clinicians (Zimmerman et al., 2004), more reflective of clinicians' personal preferences and

experiences than actual scientific evidence and/or clear cut guidelines. This situation is unsatisfactory, but a better and more evidence-based way to choose between antidepressants has yet to be identified.

One way to gain more insight into differences between antidepressants' efficacy could be to focus not only on estimated effect size, which is similar between different antidepressants and therefore not useful for differentiation, but also on the *evidence load* for this effect. The evidence load reflects the degree to which each antidepressant's efficacy is supported by the available evidence. It is important that evidence load is not confused with effect size: effect size quantifies the estimated effect of an antidepressant (e.g., the antidepressant reduces symptoms of depression by half of SD), whereas evidence load quantifies the strength of the evidence in favor of the estimated efficacy (e.g., strong evidence that the antidepressant reduces symptoms of depression). The results of such an analysis could help clinicians to choose the antidepressant with the highest evidence load for its efficacy from a range of antidepressants with comparable effect sizes. This can be done with Bayes factors (BFs) (Goodman 1999; Lavine et al., 1999; US Food and Drug Administration, 2010; Monden et al., 2016), which originate from Bayesian statistics and quantify the strength of evidence for an efficacy estimation.

To quantify the available evidence for different antidepressants' efficacy, it is important to avoid reporting bias as much as possible. Because the published literature on antidepressant efficacy has been shown to over represent positive results (Turner et al., 2008, Roest et al., 2015), data provided by the Food and Drug Administration (FDA) as part of the evaluation process may offer more conservative estimations. Trials are registered with, and results are reported to, the FDA by pharmaceutical companies to receive marketing approval. When the FDA approves a drug, the FDA reviews become publicly available, which are much less liable to the effects of reporting biases than journal articles. The current study aimed to quantify and

compare the evidence-base for the efficacy of FDA-approved second-generation antidepressants by means of BFs using data that was extracted from the FDA reviews.

Methods

Data from FDA reviews

The precise data extraction method is explained in Turner et al., (2008) and briefly summarized below. Part of the FDA reviews of second-generation antidepressants approved for Major Depressive Disorder were previously obtained (Turner et al., 2008) and an additional part for newly approved drugs (specifically levomilnacipran, vilazodone, vortioxetine and desvenlafaxine) was requested and obtained from the FDA. The phase 2/3 clinical trials for antidepressants approved by the FDA between 1987 and 2016 were identified. In total, reviews of 134 FDA-registered trials were extracted for 16 second-generation antidepressants: bupropion, citalopram, escitalopram, fluoxetine, paroxetine, paroxetine controlled release [CR], sertraline, duloxetine, mirtazapine, nefazodone, venlafaxine, venlafaxine extended release [XR], levomilnacipran, desvenlafaxine, vortioxetine and vilazodone. In line with Turner et al., (2008), the data for dosages ultimately approved by the FDA were included in the study, but the data for dosages ultimately disapproved by the FDA were excluded. From the obtained FDA reviews, the efficacy data on all randomized, double-blind, placebo-controlled studies for the short-term treatment (6-8 weeks) of depression were extracted and included in the current analyses. Ethical approval was not required for the current study as our data came from previously published studies that all received IRB approval.

Statistical Analysis

Calculation of test statistics, pooled effect sizes, and CIs

BF was calculated for each antidepressant. To do this, the sample sizes and P-values reported in the FDA reviews were used to calculate t-statistics. The derived t-statistics and the sample sizes were needed to calculate the BFs. When the precise P-value was unavailable, we estimated the t-statistics using the following three approaches, consistent with Turner et al., 2008): (a) the

mean difference score was used together with the standard deviations/standard errors/confidence intervals (CIs) to calculate precise P-values, (b) the top of the reported P-value range was used as a precise P-value (e.g., $p<0.001 \rightarrow p=0.001$), and (c) when the trial was published in agreement with the FDA conclusion (i.e., when the trial published in a journal had the same conclusion as the FDA's conclusion or when no reporting bias was found in Turner et al., 2008), the precise P-value reported in the journal was used. If (a) was impossible, we applied (b). In case both (a) and (b) were impossible, we applied (c). The above-mentioned approaches were not possible for 4 not-positive trials for paroxetine (FDA study number: 07, 09, UK-06 and UK-12) and 1 trial for sertraline (FDA study number: 310). We obtained approximations for these five trials by modeling all t-statistics/effect sizes for a given drug as coming from a truncated normal distribution. The truncation point differs for each of those trials, depending on the sample sizes used in the trial and was the point for which the associated P-value would be .05. Modeling was done using JAGS (Plummer et al., 2014), with R-package, rjags (version 4-5) (Plummer et al., 2014).

For trials with a fixed-dose design, where drug dosages were set before the trial, the t-statistic was calculated for each of the dose-levels. For a flexible-dose design, where drug dosage could be increased or remained stable over time, one t-statistic was calculated for the whole range of dosages. For trials in which placebo performed better than the study drug, the t-statistics were multiplied by -1. To compare the Bayesian and classical effect size estimations, pooled effect sizes across studies for each drug (Hedges' g) and CIs were calculated by using a random effect-pooling method, performed in STATA version 13.1 (STATA, 2013).

Bayes factors, meta-analytic Bayes factors, and Posterior effect sizes

A Bayes factor (BF_{10}), which ranges from 0 to infinity, is a ratio that quantifies the extent to which the data supports one hypothesis (H_1) over another (H_0). That is, a BF_{10} quantifies the

strength of evidence for the presence of the effect. It is important to note that BF_{10} indicates the extent to which the *existence of the drug effect* is supported by the available evidence and is not a measure of the effect size. Suppose we define H_1 as "the tested antidepressant has a positive effect on treating depression" and H_0 as "the tested antidepressant has no effect on treating depression", then, BF_{10} is the ratio between the evidence that supports the effect of the tested antidepressant and the evidence that supports the effect of the placebo. The first subscript indicates the hypothesis that is used as the numerator, H_1 in our study, and the latter subscript indicates the hypothesis that is used as the denominator, H_0 in our study. Therefore, $BF_{10}=1$ means that the data equally support H_1 and H_0, $BF_{10}>1$ means that the data support H_1 over H_0 (evidence for efficacy), and $BF_{10}<1$ means the data support H_0 over H_1 (evidence in favor of the null hypothesis). In general, a BF-range of 20-25 is often suggested to indicate "strong evidence" (Johnson, 2013).

Suppose two drugs (Drug A and Drug B) were independently tested against placebo and their BF_{10} were 20 and 0.2, respectively. Then, three conclusions can be drawn from based on these results: (1) the existence of the effect of Drug A was supported 20 times more by the evidence than the absence of the effect of Drug A (i.e., 20 over 1 = 20), (2) the absence of the effect of Drug B was supported 5 times more by the evidence than the existence of the effect of Drug B (i.e., 1 over 5 = 0.2). Note that the evidence points towards the alternative hypothesis for Drug A, but towards the null hypothesis for Drug B, and (3) the existence of the effect of Drug A was supported 100 (=20/0.2) times more by the evidence than that of Drug B. A BF_{10} can be calculated based on a result from a single trial, but when multiple trials are conducted to test the efficacy of a single drug, the overall strength of evidence can be quantified as a meta-analytic Bayes factor (meta-BF). Moreover, the overall Bayesian estimation of the effect size, i.e., the posterior distribution of the effect size, can be also obtained when calculating a meta-BF. In sum, the Bayes factor can be used to quantify evidence in favor of either the null hypothesis or

the alternative hypothesis, given the data. This is crucially different from the classical P-value, which only looks at the probability of data at least as extreme as those observed if the null hypothesis were true. As such, evidence in favor of the absence of an effect cannot be obtained with P-values.

The difference between P-values and the BF is that with more data (more trials), a BF either gets closer to infinity (stronger support for the existence of the effect) or closer to 0 (stronger support for the absence of the effect), while a P-value always becomes smaller and more likely to reject the null hypothesis when null-hypothesis is false and otherwise meander randomly and indefinitely on the 0-1 interval.

To obtain an overview and to differentiate between antidepressants with respect to the strength of evidence for the existence of the effect and the effect size, the meta-BF was calculated for each drug by using the R-package *BayesFactor* (version 0.9.10-2; Morey et al., 2015). The prior distribution (rscale) of the effect size was set to $\frac{\sqrt{2}}{2}$ (default), which follows a bell-shaped distribution with the highest probability (peak) at effect size = 0. For this study, we respected the direction of the predicted effect and excluded negative values from the prior. All t-statistics and the overall sample sizes of the drug and placebo groups for both fixed- and flexible-dose designs were combined for each antidepressant to calculate the meta-analytic BF and the posterior effect size distribution. Trials with fixed- and flexible-dose designs were combined based on the results of Khan et al. (2003), in which the authors did not find a dose-response relationship between antidepressants and symptom reduction.

Sensitivity analysis

To examine how the meta-BFs may differ based on the scale selection of the prior distribution, sensitivity analyses were conducted, by varying the scale from small ($\frac{1}{3} \times \frac{\sqrt{2}}{2}$) to large ($3 \times \frac{\sqrt{2}}{2}$). A

small scale for the prior distribution indicates a prior distribution of effect size sharply peaked around 0, which is a skeptical prior, whereas the large scale sets a prior distribution with a broader positive effect size, which is an optimistic prior. All statistical analyses, except for the calculation of the pooled effect sizes and CIs, were performed by using R version 3.2.3 (R Core Team, 2017). All the R code and data used in the present study is presented as a supplement.

Results

Meta-analytic Bayes factors

Figure 1 shows the results of meta-BFs in parentheses, together with the pooled effect sizes and CIs calculated in a classical way. To facilitate visualization, vilazodone was set as a reference and the tested antidepressants were divided into four groups according to the logarithms of the meta-BF values. **Figure 1** shows that the strength of evidence in favor of efficacy varied strongly across antidepressants. For instance, the evidence load for the efficacy, (i.e. the certainty that the drug has any positive effect given the data) of levomilnacipran was 1.87×10^{12} (=5,617,412,966,662/3) times higher than for vilazodone. The highest evidence load for efficacy was found for levomilnacipran, followed by desvenlafaxine and duloxetine. The lowest evidence load for efficacy was found for vilazodone. This indicates that all antidepressants, except bupropion and vilazodone, were found to have at least "strong evidence" for their efficacy (Johnson, 2013). **Figure 1** also shows that the estimated pooled effect sizes and CIs were relatively similar for the different drugs, making these effect sizes of limited use to differentiate between drugs.

Posterior distribution of the effect sizes

Figure 2 displays the posterior density distribution of each antidepressant's effect size estimation. The peak of the posterior distribution indicates the most probable estimation of the effect size for each antidepressant. For instance, venlafaxine shows a peak around 0.4, indicating the most probable effect size estimation lies around 0.4. In addition, **Figure 2** shows that the certainty of the effect-size estimations varies across antidepressants. For instance, the effect sizes of levomilnacipran and desvenlafaxine show a comparatively peaked distribution (reflected by the height of the distribution), indicating a higher certainty of the effect-size estimation, whereas venlafaxine XR showed a distribution with a broader density, indicating

lower certainty. In general, when trials of a given antidepressant obtained similar results, the peak of the density becomes tall, whereas the density becomes low or the distribution becomes broad if the trials of an antidepressant showed varying results. Similarly to the classical effect size estimations in **Figure 1**, the peaks of distributions in **Figure 2** lay mostly between 0.2 and 0.4. However, **Figure 2** illustrates additional differentiation between the drugs. Venlafaxine has the highest estimated effect size, followed by paroxetine and venlafaxine XR, while desvenlafaxine has the highest estimated certainty for the effect size, followed by levomilnacipran. Vilazodone has the lowest estimated effect size (around 0.1) with moderate certainty compared to the rest of the tested antidepressants.

Sensitivity analysis

The meta-BFs derived from small and large prior scales are presented in **Table 1**. With the skeptical prior (small scale) for the effect size, all the meta-BFs were higher than those with the medium scale, and meta-BFs with the medium scale were higher than those with the optimistic prior (large scale). This result is explained by the fact that the skeptical prior puts a higher expectation on effect sizes around 0, which is the case in this study, where effect sizes typically ranged from 0.2 to 0.4. Since the skeptical prior fit better to the data than the more optimistic priors, the resulting meta-BFs were highest. However, regardless of the prior scale selection, the efficacy of all FDA-approved second-generation antidepressants, except for bupropion and vilazodone, are supported by meta-BFs.

Discussion

Which antidepressant to prescribe to a depressed patient can be better chosen when both the observed effect sizes of antidepressants and the evidence load supporting each drug's effect are considered, together with other clinically relevant factors, such as drug tolerability. To gain more insight into the comparative evidence base for different antidepressants' efficacy, this study used a Bayesian framework to quantify the evidence-load for the efficacy of all second-generation antidepressants that have been approved by the FDA for the treatment of depression. The results showed that although the estimated effect sizes in Bayesian approach showed considerable overlap between drugs, the actual evidence-load for each antidepressant's efficacy (quantified by meta-BFs) varied strongly. The latter is notable given the fact that all studied antidepressants have previously been approved by the FDA as "efficacious" drugs. The presented posterior distributions of the effect sizes furthermore highlighted the differences between the antidepressants in terms of the certainty of their effect-size estimations. The estimated effect size was shown to be the highest for venlafaxine, followed by paroxetine, while levomilnacipran had the highest estimated certainty for the effect size, followed by desvenlafaxine. Vilazodone and bupropion had low effect sizes as well as low evidence for its efficacy.

The results of this study have important clinical and theoretical implications. Considering both the evidence load and effect-size certainty could be helpful for clinicians when selecting a drug. The APA guideline (2010) and several studies (Hansen et al., 2005; Gartlehner et al., 2011; Thaler et al., 2012; Linde et al., 2015) concluded that all second-generation drugs are equally efficacious to treat depression, although studies have also found different results (Lepola et al., 2003; Zimmerman et al., 2005; Cipriani et al., 2009; Consumer Reports, 2013; Kriston et al., 2014). The differences of our study and these previously published studies is that the latter performed systematic literature reviews, which have the advantage of resulting in larger

datasets than the current study, but the disadvantage that they are potentially biased by reporting bias (Turner et al., 2008; Roest et al., 2015). The current study did not look at the literature, but analyzed the FDA reviews of pre-registered trials, on which the FDA based their decision to accept a drug. By taking this approach, the current results provide a different perspective than previous reviews of antidepressant efficacy and demonstrate that the strength of evidence for different antidepressants' efficacy varies considerably, and could thus be used as an additional criterion to guide drug-prescription in clinical practice. For example, given that all the tested drugs showed strong evidence for their efficacy (i.e. BF>20), except for bupropion and vilazodone, it could be recommended that clinicians choose to prescribe venlafaxine to treat depression since this drug showed the highest effect size and was supported by the highest evidence load of all investigated drugs. This latter suggestion is in line with previous findings (Smith et al., 2002), even though antidepressants that were approved since then were also included in the current study. Of course, additional clinically relevant aspects need to be considered when prescribing antidepressants (e.g., side effects, acceptability and comorbidity), but the evidence load for an antidepressant could be a good basic criterion.

From a theoretical perspective, the current results have several implications. First, an important feature of BFs is that they can distinguish between "evidence for absence" (e.g., evidence that the effect size = 0) and "absence of evidence" (e.g., uncertain estimation of the effect size). Using the BF, the important distinction between these scenarios is possible, whereas it is not when using P-values. Second, comparisons between a classical and Bayesian way of evaluating drugs highlighted fundamental differences between the two approaches. The classical approach evaluates a trial in a dichotomous manner, typically with $P<0.05$, whereas the Bayesian approach quantifies the evidence in a continuous manner. This is important since all the tested drugs were ultimately concluded to be efficacious by the FDA, but strong variations were observed in the actual evidence load for efficacy. Problems relating to the use of

P-values as a measure of evidence have often been discussed (Goodman 1999; Ioannidis 2005; Wasserstein 2016) and the use of CIs and effect sizes has been encouraged (Wilkinson et al., 1999; Sullivan et al., 2012) to avoid making purely dichotomous decisions based on P-values. However, a recent study showed that even CIs are typically misunderstood (Hoekstra et al., 2014) and CIs cannot be used as a measure of estimation precision (Morey et al., 2015). An advantage of the currently used Bayesian approach is that it does allow for evaluation of the estimation precision, as reflected by the density distribution of the effect size certainty. Third, the results provide a comprehensive example of what can be achieved with the application of Bayesian analyses. Although it has clear advantages (Goodman, 1999), the Bayesian approach has for long been difficult to apply in practice due to limited computational speed and lack of usable software. However, thanks to the radical increase in the speed of computers over the past decade, and the development of user-friendly programs (Morey et al., 2015; Love et al., 2015), performing Bayesian analyses has become simpler and more feasible. Importantly, it has become easier to perform sensitivity analyses with various prior distributions, making sure that a result is not purely determined by one set of subjectively chosen priors (an often-heard criticism of Bayesian analyses).

The main strengths of this study lie in (1) the use of FDA registered trials, limiting the effects of reporting biases and (2) the use of a Bayesian approach to quantify strength of evidence. However, readers should also consider several study limitations. First, the differences between the estimated BFs and effect sizes across studies and/or antidepressants may partly be due to the differences in study designs (i.e., study length, initial severity) or data handling methods (i.e., missing value handling in calculation of statistics). Second, we limited our included data base as a result of reporting bias (Turner et al., 2008; Roest et al., 2015) present in this field. This choice may have resulted in exclusion of large, good quality trials that were not conducted as part of the FDA approval procedure, which could have influenced the

findings (e.g. underestimation of the evidence load for some antidepressants). Thus, the current study offers less biased results than meta-analyses based only on the published studies, but may also provide a somewhat limited view of the available data for each tested antidepressant. Third, the current study did not account for all factors that guide prescription behavior: e.g., severity (Kirsch et al., 2008; Fournier et al., 2010), side effects (FDA, 2016; Hu et al., 2004), costs (FDA, 2016) or comorbidity (Zimmerman et al., 2005). Incorporating clinically relevant factors can be done by utility analyses, weighing the evidence load and efficacy for each drug according to these factors. Also, Bayesian subgroup or meta-regression analysis could allow us to study the effect of antidepressants in more detail in future research. These studies could eventually provide a reference, which could be directly applied in clinical settings.

Conclusion

The current study shows considerable variation in the evidence-load for the efficacy of different FDA-approved second-generation antidepressants. This evidence-load could be an important criterion when choosing to prescribe a particular drug in clinical practice.

References

1. Alonso J, Angermeyer MC, Bernert S et al. Disability and quality of life impact of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatr Scand Suppl.* 2004; 420: 38-46.
2. American Psychiatric Association: Practice Guideline for the Treatment of Patients With Major Depressive Disorder. Am J Psychiatry 2010. http://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd.pdf Accessed March 4, 2016.
3. Cipriani A, Furukawa TA, Salanti G, Geddes JR et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*. 2009; 373; 746-758.
4. Cipriani A, Zhou X, Del Giovane C, Hetrick SE, Qin B, Whittington C, Coghill D, Zhang Y, Hazell P, Leucht S, Cuijpers P. Comparative efficacy and tolerability of antidepressants for major depressive disorder in children and adolescents: a network meta-analysis. The Lancet. 2016 Sep 2;388(10047):881-90.
5. Compton WM, Conway KP, Stinson FS, Grant BF. Changes in the Prevalence of Major depression and comorbid substance use disorders in the United States between 1991-1992 and 2001-2002. *Am J Psychiatry*. 2006; 163 (12):2141-2147.
6. Consumer Reports. Available from: http://www.consumerreports.org/health/resources/pdf/best-buy-drugs/Antidepressants_update.pdf (published Sep 2013, accessed 25 Oct 2015).
7. Ferrari AJ, Carlson FJ, Normal RE et al. Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. *PLOS Med*. 2013; 10(11):e1001547. doi: 10.1371/journal.pmed.1001547.

8. Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: a patient-level meta-analysis. Jama, 303(1), 47-53.
9. Gartlehner G, Hansen RA, Morgan LC et al. Comparative benefits and harms of second-generation antidepressants for treating major depressive disorder. An updated meta-analysis. Ann Intern Med 2011;155:772-785.
10. Goodman SN. Toward evidence-based medical statistics. 1: The Pvalue fallacy. (1999). Annals of Int Med. 130(12): 995-1004.
11. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med*. 1999;130;1005-1013.
12. Ioannidis JPA. Why most published research findings are false. Chance. 2005;18(4);40-47.
13. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294; 218-228.
14. Hansen RA, Gartlehner G, Lohr KN, Gaynes BN, Carey TS. Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. *Ann Intern Med*. 2005; 143; 415–426.
15. Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. Psychonomic bulletin & review, 21(5), 1157-1164.
16. Hu XH, Bull SA, Hunkeler EM et al. Incidence and duration of side effects and those rated as bothersome with selective serotonin reuptake inhibitor treatment for depression: patient report versus physician estimate. J Clin Psychiatry. 2004;65(7):959-65.
17. Jeffreys, H. (1961). Theory of probability,(Oxford: Oxford University Press).
18. Johnson VE, 2013, Revised standards for statistical evidence. PNAS, 110 (48), 19313-19317.
19. Kessler RC. The cost of depression. *Psychiatr Clin North Am*. 2012; 35(1):1-14.

20. Khan A, Khan SR, Walens G, Kolts R, Giller EL. Frequency of positive studies among fixed and flexible dose antidepressant clinical trials: an analysis of the food and drug administration summary basis of approval reports. Neuropsychopharmacology. 2003; 28 (3); 552-557.

21. Kirsch, Irving, et al. "Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration." PLoS Med 5.2 (2008): e45.

22. Kriston et al. Efficacy and acceptability of acute treatments for persistent depressive disorder: a network meta-analysis. Depress Anxiety 2014; 31: 621-630

23. Lavine M, Schervish MJ. Bayes factors: what they are and what they are not. *The American Statistician*. 1999;53(2);119-122.

24. Lepola UM, Loft H, Reines EH. Escitalopram (10–20 mg/day) is effective and well tolerated in a placebo-controlled study in depression in primary care. International clinical psychopharmacology, 2003;18(4);211-217.

25. Linde, K., Kriston, L., Rücker, G., Jamil, S., Schumann, I., Meissner, K., ... & Schneider, A. (2015). Efficacy and acceptability of pharmacological treatments for depressive disorders in primary care: systematic review and network meta-analysis. *The Annals of Family Medicine*, *13*(1), 69-79.

26. Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D. & Wagenmakers, E.-J. (2015). JASP (Version 0.7)[Computer software].

27. Monden RM, de Vos S, Morey R, Wagenmakers EJ, de Jonge P, Roest AM. (2016). Toward evidence-based medical statistics: A new look at the Bayes factor. Int J Methods in Psychiatric Research. Int J Methods Psychiatr Res, 25(4), 299-308.

28. Morey RD, Rouder JN. BayesFactor. cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf. Accessed June, 2017.

29. Morey, RD., Hoekstra, R, Rouder, JN., Lee, MD., & Wagenmakers, EJ. (2016). The fallacy of placing confidence in confidence intervals. Psychonomic bulletin & review, 23(1), 103-123.
30. Murray CJ, Vos T, Lozano R et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012; 380: 2197-223.
31. National Center for Health Statistics. Health, United States, 2010: With special feature on death and dying. Table 95. www.cdc.gov/nchs/data/hus/hus10.pdf Accessed August 11, 2015
32. National Institute for Health and Care Excellence (NICE). Depression management of depression in primary and secondary care. *London: National Institute for Health and Care Excellence*. 2007. https://www.nice.org.uk/guidance/cg023 Accessed August 8, 2015
33. National Institute of Mental Health. Major Depression Among Adults.http://www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml. Accessed September 16, 2015.
34. OECD iLibrary. http://www.oecd-ilibrary.org/sites/health_glance-2013-en/04/10/index.html?itemId=/content/chapter/health_glance-2013-41-en Accessed February 25, 2016.
35. Pratt LA, Brody DJ, Gu Q. Antidepressant use in persons aged 12 and over: United States, 2005-2008. NCHS data brief . 2011; no.76; Hyattsville MD: National Center for Health Statistics.
36. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. InProceedings of the 3rd international workshop on distributed statistical computing 2003 Mar 20 (Vol. 124, p. 125). Wien, Austria: Technische Universit at Wien.
37. Plummer M, Stukalov A. rjags. Available from: http://cran.r-project.org/web/packages/rjags/rjags.pdf; 2014.

38. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. (accessed 3 July 2017).
39. Roest AM, de Jonge P, Williams CD, de Vries YA, Schoevers RA. Reporting bias in clinical trials investigating the efficacy of second-generation antidepressants in the treatment of anxiety disorders: A report of 2 meta-analyses. *JAMA Psychiatry*. 2015;72(5):500-510.
40. Rush AJ, Prien RF. From scientific knowledge to the clinicalpractice of psychopharmacology: can the gap be bridged? *Psy-chopharmacol Bull*. 1995; 31:7–20.
41. Rush AJ, Trivedi MH, Wisniewski SR et al. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression.*N Engl J Med*. 2006; 354 ; 1231–1242.
42. Smith, D., Dempster, C., Glanville, J., Freemantle, N., & Anderson, I. (2002). Efficacy and tolerability of venlafaxine compared with selective serotonin reuptake inhibitors and other antidepressants: a meta-analysis. The British journal of psychiatry, 180(5), 396-404.
43. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP. 2013.
44. Sullivan GM, Feinn R. (2012). Using effect size-or why the P value is not enough. Journal of graduate medical education, 4(3), 279-282.
45. Thaler KJ, Morgan LC, van Noord M et al. Comparative effectiveness of second-generation antidepressants for accompanying anxiety, insomnia, and pain in depressed patients: a systematic review. Depression and Anxiety, 2012;29(6);495-505.
46. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*. 2008;358;252-260.
47. U.S. Food and Drug Administration. Guidance for the Use of Bayesian Statistics in Medical Device ClinicalTrials. http://www.fda.gov/RegulatoryInformation/Guidances/ucm071072.htm

48. U.S. Food and Drug Administration. Understanding Antidepressant medication. Available from: https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm095980.htm#3 (Accessed June 23, 2017

49. Wasserstein, RL& Lazar, NA. (2016): The ASA's statement on p-values: context, process, and purpose, The American Statistician, DOI:10.1080/00031305.2016.1154108

50. Whiteford HA, Degenhardt L, Rehm J et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*. 2013; 382:1575-86.

51. Wilkinson, L., & APA Task Force on Statistical Inference. (1999).Statistical methods in psychology journals: Guidelines and explana-tions. American Psychologist, 54,594–604.

52. Zimmerman M, Posternak M, Friedman M, Attiullah N et al. Which factors influence Psychiatrists' selection of antidepressants? Am J Psychiatry. 2004;161(7);1285-1289. http://dx.doi.org/10.1176/appi.ajp.161.7.1285

53. Zimmerman M, Posternak MA, Attiullah N et al. Why isn't bupropion the most frequently prescribed antidepressant?. The Journal of clinical psychiatry. 2005; 66(5);603-610.

Table 1 Meta-BFs with three different prior scales

Drug	Scale for the prior		
	Small ($\frac{1}{3} \times \frac{\sqrt{2}}{2}$)	Medium($\frac{\sqrt{2}}{2}$)	Large ($3 \times \frac{\sqrt{2}}{2}$)
levomilnacipran	8.00×10^{12}	5.62×10^{12}	2.14×10^{12}
desvenlafaxine	1.40×10^{12}	8.95×10^{11}	3.33×10^{11}
duloxetine	2.75×10^{10}	2.05×10^{10}	8.00×10^{9}
venlafaxine	1.40×10^{10}	1.32×10^{10}	5.62×10^{9}
paroxetine*	1.49×10^{9}	1.33×10^{9}	5.50×10^{8}
escitalopram	2.14×10^{7}	1.68×10^{7}	6.68×10^{6}
vortioxetine	2.40×10^{5}	2.02×10^{5}	8.22×10^{4}
mirtazapine	1.94×10^{4}	1.39×10^{4}	5.40×10^{3}
venlafaxine XR	1.09×10^{4}	9.13×10^{3}	3.74×10^{3}
sertraline*	5.49×10^{3}	3.35×10^{3}	1.24×10^{3}
fluoxetine	2.94×10^{3}	1.80×10^{3}	6.66×10^{2}
citalopram	2.23×10^{3}	1.46×10^{3}	5.47×10^{2}
paroxetine CR	4.56×10^{2}	3.34×10^{2}	1.31×10^{2}
nefazodone	2.08×10^{2}	1.29×10^{2}	48
bupropion	8	4	1
Vilazodone	7	3	1

Note. Drug names with * indicates that some test statistics were missing in the FDA reviews and therefore modeled and estimated by using JAGS.

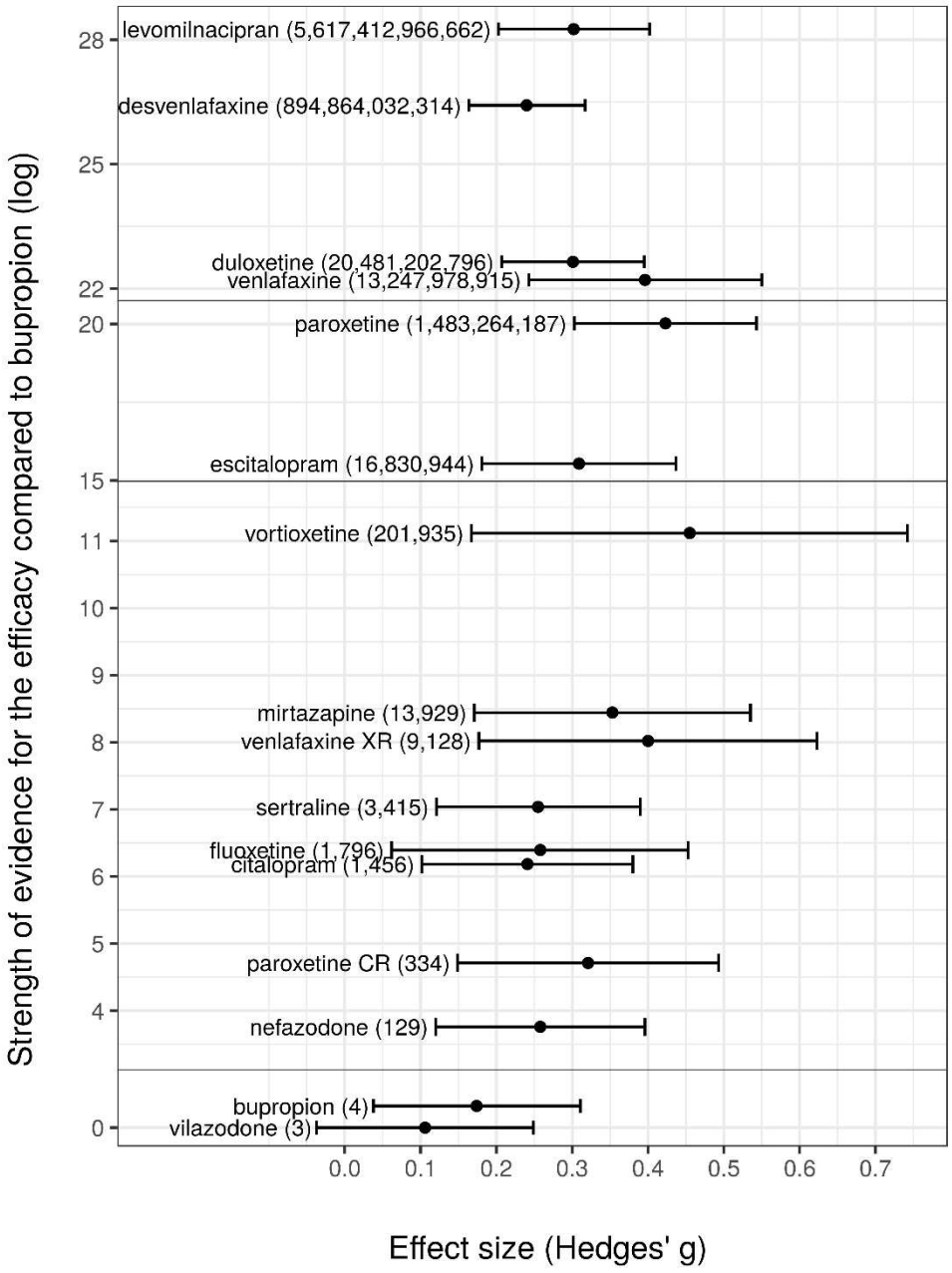

Figure 1 The relationships between Hedges'g and meta BFs

Meta-analytic Bayes factors are shown in brackets. The dots indicate effect sizes (Hedges' g). The intervals show the 95% Confidence Intervals of the effect sizes.

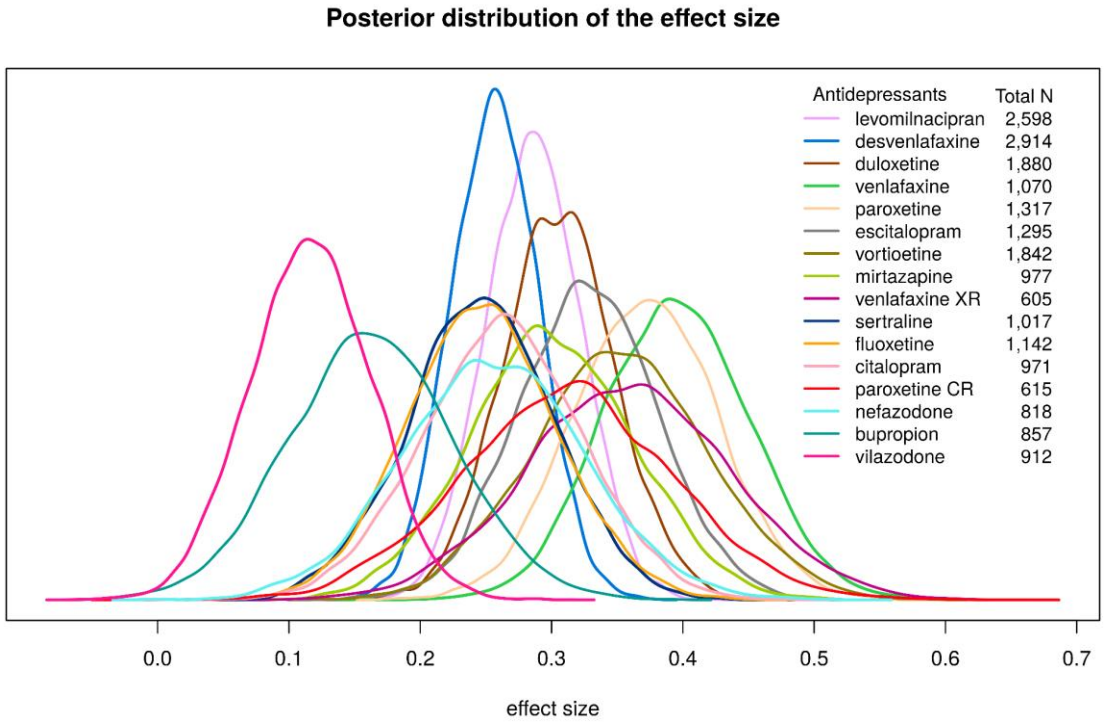

Figure 2. **Posterior effect size differences between drugs**

Density reflects the certainty of the estimated effect size and the peak of the distribution indicates the most probable estimated effect size.